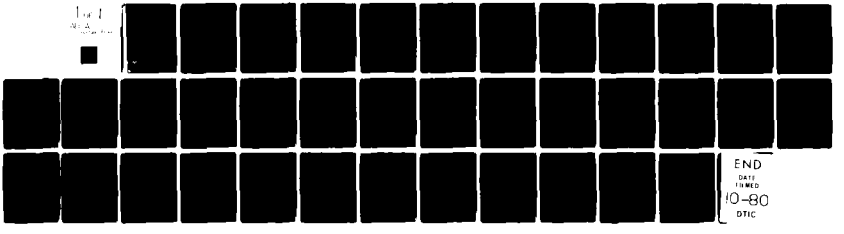
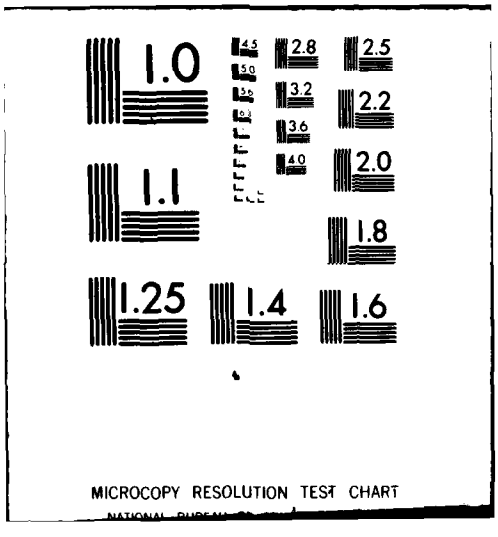
AD-A089 658 MICHIGAN UNIV ANN ARBOR COMPUTER INFORMATION AND CON--ETC F/G 12/1
SOJOURN TIMES IN MARKOV QUEUEING NETWORKS: LITTLE'S FORMULA REV--ETC(U)
JUN 80 F J BEUTLER AFOSR-76-2903

UNCLASSIFIED AFOSR-TR-80-0923 NL



END
DATE FILMED
10-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART

Sojourn Times in Markov Queueing Networks: Little's Formula Revisited

by
FREDERICK J. BEUTLER

June 1980

Research sponsored by the Air Force Office of Scientific Research, AFSC, USAF, under Grant No. AFOSR-76-2903C, and by the National Science Foundation under Grant No. ENG-75-20223.

THE UNIVERSITY OF MICHIGAN 1817
Computer Information and Control Engineering Program
Ann Arbor, Michigan 48109

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR-TR-80-0929	2. GOVT ACCESSION NO.: AD-A089 658
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): SOJOURN TIMES IN MARKOV QUEUEING NETWORKS: LITTLE'S FORMULA REVISITED	5. TYPE OF REPORT & PERIOD COVERED: Interim
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Frederick J. Beutler	8. CONTRACT OR GRANT NUMBER(s): AFOSR 76-2903 ✓
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Computer, Information & Control Engineering Program, University of Michigan, Ann Arbor, MI 48109	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: June 1980
	13. NUMBER OF PAGES: 34
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Little's formula
Queueing networks
Sojourn times
Transit times
Markov jump processes
Regenerative processes

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

It is commonly supposed that $L = \lambda W$ applies to "almost any" queueing system, with λ the mean customer entrance rate, L the asymptotic expectation of the number of customers in the system, and W the asymptotic sojourn time expectation. We study the formula for irreducible positive recurrent Markov queueing systems whose state vector Z consists of entries representing queue lengths at the respective service stations; such a model permits blocking, finite capacities, jockeying, state-dependent or random routing, bulk

UNCLASSIFIED

and/or Erlang service, and variable arrival and service rates. To define waiting times under various queueing disciplines, Z is augmented by a customer location process to yield the new Markov process Y = (Z,U). It is shown that the standard regenerative process proof of Little's equality fails in the absence of further hypotheses; however, additional assumptions assure the validity of $L = \lambda W$ for a broad variety of queueing disciplines. Generalizations of the formula are obtained for (a) state dependent customer arrival rates, (b) periodicity of the number of customers per busy period, and (c) multiple classes of customers. Non-Markovian queueing systems are briefly discussed.

(6)

SOJOURN TIMES IN MARKOV QUEUEING NETWORKS: LITTLE'S FORMULA REVISITED

(9) Interim rept.,

by

(10) Frederick J. Beutler
Computer, Information and Control Engineering Program
The University of Michigan
Ann Arbor, Michigan 48109

(11) Jun[redacted]80

(12) 37

(15) AFOSR-76-2903, NSF-ENG75-20223

Research sponsored by the Air Force Office of Scientific Research, AFSC, USAF, under Grant No. AFOSR-76-2903, and by the National Science Foundation under Grant No. ENG-75-20223.

(16) 2304

(17) A5

(18) AFOSR

(19) TR-80-0123

404621

ABSTRACT

It is commonly supposed that $L = \lambda W$ applies to "almost any" queueing system, with λ the mean customer entrance rate, L the asymptotic expectation of the number of customers in the system, and W the asymptotic sojourn time expectation. We study the formula for irreducible positive recurrent Markov queueing systems whose state vector Z consists of entries representing queue lengths at the respective service stations; such a model permits blocking, finite capacities, jockeying, state-dependent or random routing, bulk and/or Erlang service, and variable arrival and service rates. To define waiting times under various queueing disciplines, Z is augmented by a customer location process to yield the new Markov process $Y = (Z,U)$. It is shown that the standard regenerative process proof of Little's equality fails in the absence of further hypotheses; however, additional assumptions assure the validity of $L = \lambda W$ for a broad variety of queueing disciplines. Generalizations of the formula are obtained for (a) state dependent customer arrival rates, (b) periodicity of the number of customers per busy period, and (c) multiple classes of customers. Non-Markovian queueing systems are briefly discussed.

Key Words: Little's formula, queueing networks, sojourn times, transit times, Markov jump processes, regenerative processes

0. INTRODUCTION

According to a recent survey, the "... average waiting and response times can be obtained easily from Little's formula $L = \lambda W$ ..." for customers in Markovian queueing systems of general type [14]. This would seem to be supported by the classical queueing literature [12][18], in which the reference is to queueing systems rather than one server. Although these assessments may be correct in principle, a closer examination of the applicability of the formula discloses some subtleties, which do not occur in a G/G/1 queue, but may well be encountered in a complex queueing network. These considerations affect not only the form, but even the validity of the relation $L = \lambda W$. Among the deviations from G/G/1 behavior that preclude a straightforward argument are variable entry rates, batch inputs, periodicity in the number of customers in a busy cycle, and some (likely pathological) non-existent expectations.

Little's formula is amenable to diverse interpretations (e.g., as a relation for time averages [18]), but we shall confine ourselves to asymptotic expectations. If $L(t)$, the total number in the system at time t, converges in distribution to L_∞, we call $L = E(L_\infty)$. Analogously, when W_n, the system sojourn time of the n'th customer, converges in distribution to W_∞, we let $W = E(W_\infty)$. Finally, λ is the average rate of customer entry into the system.

In common with many other studies on queueing networks [1] [2][3][11][13][16], we shall represent the state of the network by a Markov jump process Z for positive time. Roughly speaking, $Z(t)$ is a complete descriptor of the location and type of customer

within the network at time t. As the references just listed indicate, the Markov representation is adequate to model a variety of computer-communication systmes, including different queueing disciplines, random routings, blocking, state-dependent entry and service rates (e.g., finite buffers and multiple servers, respectively), non-exponential service time distributions, different classes of customers, and the like. The reader is referred to the above references for applications and examples.

That the relation $L = \lambda W$ may fail even for a simple Markov system is indicated by

Example 0.1: The system consists of two exponential tandem-connected servers, with a Poisson input of intensity λ to the first server. The state is then described by a tuple $z = (z_1, z_2)$. Now suppose blocking and finite capacity permit only certain direct transitions, viz. $(0,0) \to (1,0) \to (2,0) \to (1,1) \to (0,2) \to (0,1) \to (0,0)$. Since the W_n have distinct expectations according as n is odd or even, the sequence W_n cannot converge in distribution, nor is there even any meaning to a limit of $E(W_n)$. Further, the arrival rate to the system may well be constant at λ, but this does not reflect the actual entry rate, which is some variable quantity. Consequently $L = \lambda W$ makes no sense, nor is it obvious that there is a "fixed up" version of this formula.

This example illustrates the folly of an automatic assumption that $L = \lambda W$ must hold for a stable Markov system, Apparently, variable entry rates and busy cycle periodicity play some role in the validity of the formula. We shall also see that there are other hypotheses which we cannot omit.

We shall begin by summarizing standard arguments leading to Little's formula. Next, we examine Markov queueing networks to determine their conformance to the hypotheses underlying a proof of Little's formula. Some of these hypotheses are automatically met by any queueing system represented by a well-behaved Markov process, whereas others require separate assumptions, or lead to a modification of $L = \lambda W$. Among the extensions are variable input rates and multiple customer classes. Finally, we point out that certain of the results are applicable to non-Markov networks.

I. BACKGROUND

A Markov queueing network will be represented by a vector Markov jump process defined on a state space whose elements $z = (z_1, z_2, \ldots, z_r)$ have non-negative integer-valued components z_i that may be thought of as the respective number of customers located at the r service stations; this is consistent with many literature descriptions of computer-communication system models [1][3][9][14]. Accordingly, we shall begin with a regular right-continuous Markov jump process $Z = (Z_1, Z_2, \ldots, Z_r)$ having this state space, and defined for times $t \geq 0$. The state $\theta = (0,0,\ldots,0)$, corresponding to an empty system, is supposed positive recurrent. Since only asymptotic properties are of interest, we shall take $Z(0) = \theta$ without loss of generality. Central to the upcoming derivations are the stopping (hitting) times $\{C_n\}$; $C_o = 0$, and C_n is the time of the n'th entrance of the system into the state θ. At this juncture, we only note that $\{C_n\}$ constitutes a renewal process with finite inter-renewal periods satisfying

$E(C_{n+1}-C_n) = \pi_\theta^{-1} > 0$ (cf. [5], Section 8.5 and Prop. 10.1.12), in which π_θ denotes the invariant probability measure $\lim_t P[Z(t)=\theta] = \pi_\theta$.

For the present, let us understand the sojourn times $\{W_n\}$ and number of customers in the system $\{L(t)\}$ in intuitive terms; in any case, $\{W_n\}$ cannot be defined entirely in terms of Z, since individual sojourn times are also heavily dependent on the queueing discipline (not reflected in Z). For now, we review a derivation of Little's formula for a later analysis of its application to a queueing network specified by the Markov process Z. Of the modern proofs of Little's equality, perhaps the best known and most general are those of Jewell [12] and Stidham [18] [19]. Stidham points out that the relation, in terms of time averages, is actually non-stochastic; under further assumptions, $L = \lambda W$ can be retrieved for a random system. Jewell's approach (see also [6], Section 1.6ii) is more direct, attaining its results through use of regenerative process theory and Wald's lemma. Unfortunately, Jewell ignores the implications of periodicity and non-uniform arrival rates. It is easy to state the propositions underlying Jewell's argument, and we do so for future reference without proof.

Theorem 1.1 (cf. [18], Theorem I; [12], Lemma 1): For a busy cycle over the interval $[0,C]$, and entailing N customers

$$\int_0^C L(t)dt = \sum_{n=1}^{N} W_n \quad , \tag{1.1}$$

where W_n is the total system sojourn time for the n'th customer

to enter the system, and $L(t) = \| Z(t) \|$ (i.e., the ℓ_1 norm of $Z(t)$).

Theorem 1.2 ([6], Theorem 5.1): Let Z be a regenerative process ([6], Definition 2.1) with respect to the renewal sequence $\{C_n\}$, where the renewal intervals are non-arithmetic with finite expectation. If the expectation on the right of (1.2) exists, $L(t)$ converges in distribution to a random variable L_∞, and

$$E(L_\infty) = \{E(C)\}^{-1} E\{\int_0^C L(t)dt\} \quad . \tag{1.2}$$

Remarks: Without loss of generality, we have taken $C_o = 0$, and used C for C_1 in (1.2). Although it does not affect the statement of the Theorem, we prefer to let the C_n correspond to times the system becomes empty, rather than--as convention would have it--those instants the empty system receives a first customer. Our definition assures that Z is in the specific state $Z(C_n) = \theta$, so that we may use the strong Markov property.

Theorem 1.3 ([6], Theorem 3.2): Let $\{W_n\}$ be a regenerative process with respect to the renewal sequence $\{N_k\}$, were N_k is the total number of customers entering the system over the interval $(0, C_k]$. If the number of customers in a busy period has finite expectation and is aperiodic, the W_n converge in distribution to a random variable W_∞, and

$$E(W_\infty) = \{E(N)\}^{-1} E\{\sum_{n=1}^{N} W_n\} \quad . \tag{1.3}$$

Remark: As in Theorem 1.1, we take $C_o = 0$ and write N for N_1.

Theorem 1.4 (Wald's Lemma): Assume $\{A_n\}$ is a sequence of i.i.d. random variables of finite mean. Let N be a random variable such that the event $\{N \leq n\}$ is stochastically independent of A_k for all $k \geq n+1$. Then if C is defined as

$$C = \sum_{n=1}^{N} A_n \tag{1.4}$$

we have

$$E(C) = E(A)E(N) \quad . \tag{1.5}$$

Remark: If the busy cycle is conventionally defined to begin with the arrival of a customer into the empty system, the A_k are interarrival times, and N is as in Theorem 1.3, then (1.4) correctly describes the busy cycle (cf. [6], p.1); hence (1.5) holds for i.i.d. interarrival periods at the constant rate $\lambda = [E(A)]^{-1}$. However, (1.4) is not appropriate for our busy cycles $(C_n, C_{n+1}]$, which are assumed to begin when the system first becomes empty.

If the conditions of the preceding theorems are satisfied, we may combine equations (1.1), (1.2), (1.3) and (1.5) to yield Little's formula with $L = E(L_\infty)$ and $W = E(W_\infty)$. As is well known, the formula is applicable to the stable G/G/1 queue--provided we ignore periodicities. For instance, the special case of the D/G/1 queue (e.g., $A_n = 1$) implies a periodic L(t) that does not converge in distribution. We also find that an adjustment must be made for fixed bulk arrivals, and that Theorem 1.4 fails altogether if the G/G/1 queue is modified by introducing state dependence in the arrival process. Although arrival periodicity is inconsistent with the Markov property as it appears in a

Markov queueing network, we cannot exclude stochastic dependence of the state, the service process, and the arrival process. These complex behavior modes suggest that Little's relation may not be generally applicable, as Example 0.1 has already indicated. We are therefore led to a careful construction and examination of a Markov network queueing model, for which we ascertain the extent to which the hypotheses of the above theorems are satisfied. When these theorems are not directly applicable, we inquire whether the Markov model must meet additional restrictions, or whether the original model can lead to modified forms of Theorem 1.1 to 1.4.

2. APPLICABLE PROPERTIES OF MARKOV QUEUEING NETWORKS

Because the queueing network state Z is a Markov process, the hypotheses of Theorems 1.1 to 1.4 are met in varying degrees. A portion of this section is devoted to reviewing these assumptions in light of the Markov character of Z, and determining whether additonal suppositions are needed for a proof of Little's formula. We see, however, that Z lacks altogether the customer identification essential to tracing a customer's progress through the system (for any queueing discipline), and relating his emergence from the system to his entry; hence, it is impossible to define the sojourn times W_n without further structure. Augmentation of Z to a new Markov jump process will prove useful in this regard, eventually leading to forms of Little's equation that are robust respective to queueing disciplines.

We recall that Z is a regular right-continuous Markov jump

process for which θ is a positive recurrent state; to avoid trivialities, we require that θ not be absorbing. We can now characterize the Markov process Z completely in terms of the times of jump T_n, and the values of Z on its intervals of constancy. If $\{Z^n\}$ is defined by

$$Z^n = Z(T_n) \quad , \tag{2.1}$$

$\{Z^n\}$ and the tuple $\{(Z^n, T_n)\}$ are both Markov processes, as is evidenced by (see [5], Chapter 8, Section 3)

$$P[Z^{n+1}=z, T_{n+1}-T_n>t \mid Z^n=x, \Lambda] = q_{xz}\exp[-\alpha(x)t] \quad . \tag{2.2}$$

Here Λ is any event in the sigma algebra generated by $\{Z^0,\ldots,Z^n, T_0,\ldots,T_n\}$, and $0 < \alpha(x) < \infty$. The q_{xz} are transition matrix elements for the discrete Markov process $\{Z^n\}$. Further, (2.2) is related to the infinitesmal operator matrix for Z in the sense that the matrix entries (sometimes interpreted as flow intensities) are $a_{xz} = \alpha(x)q_{xz}$ unless $x = z$, in which case $a_{xx} = -\alpha(x)$.

Because we demand that Z be a queueing process, we restrict possible positive q_{xz} to those consistent with queueing models. Thus, the transition probability $q_{xz} = 0$ except (perhaps) in the following three situations:

k customers enter the system at service station i, which means

$$z = x+ke_i \quad ; \tag{2.3}$$

(e_i is the unit vector whose i'th component is unity); k customers depart from station i and exit the system, i.e.,

$$z = x-ke_i \quad ; \tag{2.4}$$

or k customers transfer from service station i to arrive immediately at j, viz.

$$z = x - ke_i + ke_j \quad . \tag{2.5}$$

While the q_{zk} and $\{Z^n\}$ specify queue lengths and their variation, they fail to reveal the location, routing and departure of individual customers. To speak of sojourn times, therefore, we must introduce an auxiliary process, the queueing discipline process $\{U^n\}$. For the empty system (i.e., $Z^n = \theta$), $U^n = \theta$. Within each busy cycle, U^n includes for each customer a triple (m,q_m,j_m), in which m is the customer's order of arrival in chronological order, q_m (with $0 \leq q_m \leq r$) the queueing station to which he is assigned, and j_m (with $0 \leq j_m \leq z^n_{q_m}$) his position at station q_m. The entries $q_m = j_m = 0$ are reserved for customers m who have already left the system. Of course, U^n is required to be consistent with Z^n. Thus, a transition (2.3) introduces new triplets to U^n, while a departure (2.4) reduces k appropriate pairs q_m and j_m to zero. Along with any type of transition, there may also be a permutation of customer positions.

It is essential that the queueing discipline process be of the functional form

$$U^{n+1} = \psi(U^n, Z^n, Z^{n+1}, \xi^{n+1}) \tag{2.6}$$

where the ξ^n are i.i.d. variates. A great variety of queueing disciplines can be represented in this fashion. These include FCFS, LCFS, random service order, and even state-dependent disciplines. However, we cannot admit discipline dependencies that reach across different busy cycles, or disciplines that depend on states prior to those indicated by (2.6).

By virtue of (2.6), $\{(Z^n,U^n)\}$ is a Markov chain, which in turn generates to continuous parameter Markov process Y, where

$$Y(t) = (Z^n,U^n) \quad \text{for} \quad T_n \leq t < T_{n+1} \quad ; \qquad (2.7)$$

see again [5], Section 8.3 for details. We note for future reference that if C_k is the hitting time at which $Z(C_k)$ becomes null for the k'th time C_k assumes the same role for Y. The new process Y is used to define sojourn times. First, let $I_{k,m}(t) = 1$ if $C_k \leq t < C_{k+1}$ and if Y includes the m'th customer through the triple (m,q_m,j_m) with nonzero q_m and j_m; otherwise, $I_{k,m}(t) = 0$. With the aid of this indicator function we can define unambiguously the customer sojourn times by

$$W_{m+N_k} = \int_{C_k}^{C_{k+1}} I_{k,m}(t)dt \quad . \qquad (2.8)$$

Here N_k is the total number of customers in the first k busy cycles. More precisely, the number of arrivals in (0,t] is

$$M(t) = \sum_{T_i \leq t} [\,\| Z(T_i) \| - \| Z(T_i-) \|\,]^+ \qquad (2.9)$$

and $N_k = M(C_k)$. This definition of N_k is consistent in the sense that $(N_k - N_{k-1})$ arrivals are indicated by U over $(C_{k-1},C_k]$.

Let us now turn to the contents of Theorem 1.1. We state and prove

Theorem 2.1: If $\{U^n\}$ is consistent with $\{Z^n\}$, the assertion (1.1) of Theorem 1.1 is valid for the queueing system Y.

Proof: At $t = 0$, we have $L(0) = 0$ as well as $I_{0,m}(0) = 0$ for all n. For each jump in $L(t) = \| Z(t) \|$, there is a corresponding jump

in the sum of the $I_{0,m}(t)$, that is

$$L(t) = \sum_{m=1}^{\infty} I_{0,m}(t) \quad . \tag{2.10}$$

On the right side of (2.10), the terms are identically zero for $m > N_1$, so that the sum is adjusted accordingly. If we now integrate both sides of (2.10) on $(0,C_1]$, we obtain the desired result from (2.8).

Remark: We again note that (1.1) is "discipline robust" with respect to any queueing discipline obeying the consistency condition.

We have seen in Section I that regenerative processes are central to Jewell's proof of Little's formula. We now connect Markov queueing networks, as described by Y, with regenerative process theory. Most useful for this purpose is the strong Markov property, whose relevant parts we present below; a proof can be found in [4], Section II.9, Theorem 5. We must again begin with certain definitions that make precise the "past" and "future" with respect to the C_n. To begin with, let $\mathcal{F}_t = \sigma\{Y(s), s \leq t\}$, that is, $\mathcal{F}_t$ is the σ-algebra generated by Y up to time t. Next, the pre-T σ-algebra $\mathcal{F}_T$ is specified in standard fashion for a stopping time T as follows: $\Lambda \in \mathcal{F}_T$ if $\Lambda \cap \{T \leq t\} \in \mathcal{F}_t$ for each t. We also define a post-T ("future") σ-algebra $\mathcal{F}^T$. Let S_T be the translation operator $(S_T X)(t) = X(t+T)$; then $\mathcal{F}^T = \sigma\{(S_T Y)(t), t \geq 0\}$. With this notation, we state

Lemma 2.2: Let $0 = C_0 < C_1 < C_2 < \ldots$ be a sequence of stopping times such that $P[Y(C_k) = \theta] = 1$ for $k = 0,1,2,\ldots$ If $\Lambda \in \mathcal{F}_{C_k}$ and

$\Gamma \in \mathcal{F}^{C_k}$, then

$$P(\Lambda \cap \Gamma) = P(\Lambda)P(\Gamma) \quad . \tag{2.11}$$

Moreover, probabilities are invariant under translations S_{C_k};

$$P[\left(S_{C_k}(Y(t_1),Y(t_2),..,Y(t_m))\right) \in B] = P[(Y(t_1),Y(t_2),..,Y(t_m)) \in B]. \tag{2.12}$$

Remark: Lemma 2.2 would fail to hold if we had taken the C_k as arrival times of the first customer into the empty system, unless $q_{\theta z} > 0$ for exactly one z, implying that the first arrival number and location must be fixed. To avoid such a restriction we have defined the C_k as already indicated, contrary to the usual convention of the initiation of the busy cycle. With the aid of Lemma 2.2, it is easy to prove

Theorem 2.3: Theorem 1.2 holds for the queueing system Y.

Proof: It was noted near the beginning of Section I that $\{C_n\}$ is a non-arithmetic renewal process whose intervals have finite expectation. From (2.11) we are able to conclude that the sigma algebras $\sigma\{L(t), t \leq C_k\}$ and $\sigma\{L(t), t > C_k\}$ are independent for each k. Lastly, every $S_{C_k} L$ has the same probability distribution. In other words, L is properly regenerative with respect to $\{C_n\}$.

Remark: More generally, of course, Y is regenerative with respect to $\{C_n\}$.

The result pertinent to Theorem 1.3 is, unfortunately less complete:

Theorem 2.4: $\{W_n\}$ is regenerative with respect to the renewal sequence $\{N_k\}$.

Proof: Since $N_k = N(C_k)$ (see (2.9)), N_k is $\mathcal{F}_{C_k}$ measurable, and $N_{k+m} - N_k$ is $\mathcal{F}^{C_k}$ measurable; hence, the intervals $(N_{k+1} - N_k)$

are mutually independent according to (2.11). That they are also identically distributed is a consequence of (2.12). Moreover, N_1 has an honest distribution, or equivalently, $P[N_1 < \infty] = 1$. To prove this, return to the discrete parameter process $\{Z^n\}$, letting c_k be the k'th hitting epoch for state θ. Since θ is positive recurrent for Z, θ is likwise positive recurrent for $\{Z^n\}$. Then c_1 is almost surely finite. If we now rewrite (2.9) in the form

$$N_1 = \sum_{n=1}^{c_1} [\,\|Z^n\| - \|Z^{n-1}\|\,]^+ \qquad (2.13)$$

and note that each summand must be finite, we have $P[N_1 < \infty] = 1$.

It remains to show that $W_1, W_2, \ldots$ have the same multivariate distribution as $W_{1+N_k}, W_{2+N_k}, \ldots$ for every k. Since $S_{C_k} I_{0,n} = I_{k,n}$ it follows from (2.12) that the joint distribution of the $I_{0,n}$ coincides with that of the $I_{k,n}$. The definition (2.8) of W_{n+N_k} then leads to the desired result.

Corollary 2.5: Almost surely

$$\sum_{n=1}^{N_1} W_n < \infty \quad . \qquad (2.14)$$

Proof: For each $n = 1,2,\ldots,N_1$, we have $W_n < C_1$. Since $P[C_1 < \infty] = 1$ from $E(C_1) < \infty$, the proof is complete.

Although both sides of (1.1) are seen to be finite from the corollary just stated, more is needed. None of our assumptions preclude any of

$$E(N) = \infty \;, \quad E[\sum_{n=1}^{N} W_n] = \infty \;, \quad E[\int_0^C L(t)dt] = \infty \quad , \qquad (2.15)$$

in which we have again written N for N_1 and C for C_1. The first expression in (2.15) is contrary to a hypothesis in Theorem 1.3, in which finiteness is needed to make sense of (1.3) and (1.5). The second and third expectations in (2.15) are finite or infinite together, as may be seen from Theorem 2.1. To illustrate (2.15), we offer an example involving a single Markov server.

Example 2.6: From (2.2), the single Markov server is described by $\{q_{ij}\}$ and $\{\alpha(i)\}$. We take $q_{0k_1} = 1$ and with $0 < k_1 < k_2 < ..$ let $q_{k_n k_{n+1}} = p$, $q_{k_n 0} = 1-p$. Then

$$E(C) = p^{-1} \sum_{n=1}^{\infty} [\alpha(k_n)]^{-1} p^n \quad , \tag{2.16}$$

so that Z is recurrent for any $p < 1$ provided that $\underline{\lim}_n [\sqrt[n]{\alpha(k_n)}] \geq 1$; this is an easy consequence of the theory of power series ([10], Section 5.4). We further find that

$$E(N) = (1-p)p^{-1} \sum_{n=1}^{\infty} k_n p^n \quad . \tag{2.17}$$

It is apparent that the free choice of the rate of growth of $\{k_n\}$ allows on the one hand the possibility that $E(N) < \infty$ for any $p < 1$, or on the other (by taking $\overline{\lim}_n [\sqrt[n]{k_n}] = \infty$) that $E(N) = \infty$ for every $p > 0$. Finally, we calculate

$$E[\int_0^C L(t)dt] = p^{-1} \sum_{n=1}^{\infty} k_n [\alpha(k_n)]^{-1} p^n \quad . \tag{2.18}$$

Both finite and infinite expectations in (2.17) and (2.18) are compatible with $E(C) < \infty$ in any combination, as we shall show.

(a) Let $\{k_n\}$ satisfy $1 \leq \underline{\lim}_n [\sqrt[n]{k_n}] \leq \overline{\lim}_n [\sqrt[n]{k_n}] = p_o^{-1} < \infty$, and take $\alpha(k_n) = k_n$. Then $E(C) < \infty$ and

$$E[\int_0^C L(t)dt] < \infty \tag{2.19}$$

for any $p < 1$, while $E(N) = \infty$ for $p > p_o$ and $E(N) < \infty$ for $p < p_o$.

(b) If $\alpha(k_n) = 1$ for all n, $E(C) < \infty$ for any $p < 1$. Here (2.10) is true iff $E(N) < \infty$.

(c) We choose $\{k_n\}$ and $\{\alpha(k_n)\}$ such that $E(C) < \infty$, $E(N) < \infty$, but

$$E[\int_0^C L(t)dt] = \infty \tag{2.20}$$

Take $\underline{\lim}_n [\sqrt[n]{\alpha(k_n)}] = \frac{2}{3}$, and $\overline{\lim}_n [\sqrt[n]{k_n}] = 2$; the radii of covergence of the power series (2.16), (2.17) and (2.18) are then respectively $\frac{2}{3}$, $\frac{1}{2}$, and $\frac{1}{3}$, leading to different possibilities that depend on the selection of p. Specifically, for $\frac{1}{3} < p < \frac{1}{2}$ we obtain the asserted expectations.

One might infer from the example that some of the anomalies in infinite expectation are connected with the possibility of customer entry in arbitrarily large batches. Accordingly, we study the behavior of these expectations under constraints on the number of customers that can enter at any one time.

Theorem 2.7: Suppose $q_{xz} = 0$ whenever $z = x+ke_i$ for $k > k_o$, where k_o is a fixed integer. Then $E(N) < \infty$.

Proof: From (2.13), $N_1 < k_o c_1$ and, since the positive recurrence of Z implies that of $\{Z^n\}$, $E(c_1) < \infty$.

Remark: There is no point in analyzing systems with $E(C) = \infty$, since then L and W are both infinite.

It is tempting to conclude from Theorem 2.7 that bounded inputs yield also the finiteness of $E[\int_0^C L(t)dt]$. Unfortunately, such an assertion is false in the absence of additional hypotheses, so that the finiteness of L and W cannot be taken for granted. In the following illustration, we see that (2.20) can apply even for the single server Markov queue with individual arrivals and services.

Example 2.8: The discrete parameter embedded Markov chain is a simple random walk (birth-and-death process) with $q_{01} = 1$, $q_{10} = \frac{1}{2}$, $q_{12} = \frac{1}{2}$, and for $n \geq 2$, $q_{n,n+1} = \frac{1}{2}(\frac{n+1}{n})$ and $q_{n,n-1} = \frac{1}{2}(\frac{n-1}{n})$. If we take $\alpha(n) = 1$ for all n, it follows that the relevant expectations can be written in terms of the discrete parameter process $\{Z^n\}$, that is, $E(C) = E(c)$ and

$$E[\int_0^C L(t)dt] = E[\sum_{n=1}^{c} Z^n] \quad . \tag{2.21}$$

Following [4], Section I.12 and I.14, one calculates (ignoring periodicity) that $E(c) < \infty$, and that for $n \geq 2$ the stationary probabilities satisfy $\nu_n = 2[(n+1)(n-1)]^{-1}\nu_o$. Moreover,

$$E[\sum_{n=1}^{c} Z^n] = \nu_o^{-1}\sum_{n=1}^{\infty} n\nu_n = \infty \quad . \tag{2.22}$$

The above result remains qualitatively unchanged for other $\{\alpha(n)\}$ provided only that $0 < \inf[\alpha(n)]$ and $\sup[\alpha(n)] < \infty$. Thus, there

appear to be reasonable Markov queueing models for which $E(L_\infty)$ and $E(W_\infty)$ are both infinite.

In view of Theorems 2.4 and 2.7, Theorem 1.3 is valid under the additonal assumption that N is aperiodic. Nevertheless, unless L and W are finite, we cannot arrive at Little's formula. We must therefore seek conditions under which

$$E[\int_0^C L(t)dt] < \infty \quad . \tag{2.23}$$

We now--and hereafter--suppose that only one customer can enter at a time (i.e., $k_o = 0$). We also assume (and shall shortly discuss in detail) that the interarrival times A_k mentioned in Theorem 1.4 are indeed i.i.d. random variables. Then we shall have

Theorem 2.9: Let $E(N^2) < \infty$. Then (2.23) holds.

Proof: We will majorize the integral $\int_0^C L(t)dt$ as follows. Let $V_j = \inf\{t:M(t)=j\}$, so that $A_j = V_j - V_{j-1}$. Since $C < V_{N+1}$ and $L(t) \leq j$ on $(V_j, V_{j+1}]$ we obtain

$$\int_0^C L(t)dt < \sum_{j=1}^{N} jA_{j+1} < \sum_{j=1}^{N+1} jA_j = \sum_{j=1}^{\infty} jA_j I_{\{j \leq N+1\}} \ , \tag{2.24}$$

in which I is a self-evident indicator function. On the right of (2.24), the event $\{N \leq j-2\}$ must be independent of A_j (an interarrival time belonging to the next busy cycle). Consequently, the expectation of the right side of (2.24) simplifies; with $E(A_j) = \lambda^{-1}$, we have

$$E[\int_0^C L(t)dt] < \lambda^{-1}[E(N^2)+E(N)] \quad . \tag{2.25}$$

Remark: It may be inconvenient to verify whether $E(N^2)$ is finite. However, $N < c$, and the recurrence time c is completely described in terms of the transition matrix Q for $\{Z^n\}$; see [4], Section I.11 for specific properties of $E(c^2)$. We also note that there is no need to assume that the A_j have finite expectation. In fact, A_1 (and hence every A_j) is exponentially distributed with parameter $\alpha(\theta)$.

So far, we have studied the applicability of Theorems 1.1 through 1.3 for Markov queueing systems. Now we turn to the substance of Theorem 1.4. As we have noted above, A_1 has an exponential distribution with parameter $\alpha(\theta)$, so that $\{A_k\}$ is a Poisson process if the variates are i.i.d. in accordance with the hypotheses of Theorem 1.4. It is essential to realize that the interarrival intervals A_k are wholly expressed within the structure of Z; the arrival process M is incremented whenever the state z jumps to $z+e_i$ for some $i = 1,2,..,r$. However, even for a Poisson process K induced by a Markov Z, intervals such as those described in Theorem 1.4 need not be i.i.d. when the indices are stopping times for $\mathcal{F}_t$ instead of the smaller sigma algebra $\mathcal{M}_t = \sigma\{M(u), u \leq t\}$. A simple example illustrates this phenomenon. Let $\{D_k\}$ be the interdeparture intervals associated with the departure process K of a M/M/1 queue in equilibrium, and take C as the time when the system first becomes empty. It is known that K is Poisson; nevertheless, $D_{K(C)} = \inf\{t: K(t)-K(C)=1\} - C$ is neither exponentially distributed, nor independent of $D_{K(C)+1}$. Thus, we must assure that the A_k are i.i.d. by an assumption that goes beyond the independence of $[M(t)-M(s)]$ from $\mathcal{M}_s$ to permit

the application of $\{\mathcal{F}_t\}$ stopping times. Specifically, the evolution of the arrival process must be independent of the current state of the queueing system, as is actually true whenever the customer stream is not influenced by internal system conditions.

We formalize the above discussion through

Hypothesis 2.10: $\lim_{h \searrow 0} h^{-1} P(M(s+h)-M(s)=1 | Z(s)] = \lambda$ (2.26)

for all $s \geq 0$.

Evidently, Hypothesis 2.10 states that the rate of accession of customers into the system depends neither on the time nor on the state of the system. The Hypothesis provides some of the prerequisites for Theorem 1.4, as is indicated by

Theorem 2.11: Hypothesis 2.10 is equivalent to

$$P[M(t)-M(s)=n | \mathcal{F}_s] = \frac{[\lambda(t-s)]^n}{n!} e^{-\lambda(t-s)} \qquad (2.27)$$

for every n and $t \geq s \geq 0$. If one of these conditions is satisfied, M is a Poisson process, and the A_k are i.i.d. random variables.[1]

Proof: In the presence of (2.26), the forward Kolmogorov equation for the Markov process Z becomes

$$\frac{d}{dt}\{P[M(t)-M(s)=n | Z(s)=z]\} = \lambda\{P[M(t)-M(s)=n-1 | Z(s)=z] - P[M(t)-M(s)=n | Z(s)=z]\} \qquad (2.28)$$

where $\{M(t)-M(s)=-1\}$ has zero probability, and an initial

---

[1] If the conditioning in (2.26) is on $\mathcal{F}_s$, M is already a Poisson process respective to $\{\mathcal{F}_t\}$, without reference to the Markov nature of Z. One notes that the compensator of the submartingale M is given by [15], VII.T28 and VII.T29, which is almost surely equal to λt by comparison with (2.26). The desired result then follows from Watanabe's theorem ([5], p. 76).

condition is furnished by $P[M(s)-M(s)=0] = 1$. If we now solve the system (2.28) recursively, we obtain (2.27), except that the conditioning is on $Z(s)$. However Z is Markov, so that the conditioning on $Z(s)$ may be replaced by the sigma algebra $\mathcal{F}_s$.

To show that M is Poisson, take the conditional expectation of (2.27) with respect to $\mathcal{M}_s = \sigma\{M(u), u \leq s\}$. Since $\mathcal{M}_s \subset \mathcal{F}_s$, (2.27) remains true if the conditioning is on $\mathcal{M}_s$ rather than with respect to $\mathcal{F}_s$. Next, we consider the A_k. As before, we define $V_k = \inf\{t: M(t)=k\}$. Then $\sigma\{A_j, j \leq k\} \subset \mathcal{F}_{V_k}$, and we show that A_{k+1} is exponentially distributed and independent of $\mathcal{F}_{V_k}$. In fact,

$$P[A_{k+1} > a \mid \mathcal{F}_{V_k}] = P[M(V_k+a)-M(V_k)=0 \mid \mathcal{F}_{V_k}] \quad ; \qquad (2.29)$$

by the strong Markov property, and because of (2.28), the right side of (2.29) is simply $e^{-\lambda a}$. This demonstrates both the independence and exponential distribution of the A_k.

Lastly, (2.27) is readily derived from (2.28), so the proof of the Theorem is complete.

Corollary 2.12: $\lambda = \alpha(\theta)$ (2.30)

Proof: In (2.27), let $n = 0$ and $s = 0$. Since $\{M(t)=0\} = \{T_1 > t\}$, we can compare (2.27) with $P(T_1 > t)$ as obtained from (2.2).

We could now show the applicability of Theorem 1.4 when Hypothesis 2.10 holds. Although the length of the first busy cycle, as defined herein, does not satisfy (1.4), one can show that

$$E(C) = E[\sum_{1}^{N+1} A_k] - E(A_1) \quad , \qquad (2.31)$$

from which (1.5) eventually follows. However, Hypothesis 2.10

permits us to proceed more directly via a simpler proof, which moreover requires no recourse to Wald's Lemma.

Theorem 2.13: Under Hypothesis 2.10, we have

$$E(C) = \lambda^{-1} E(N) \quad . \tag{2.32}$$

Proof: The process

$$K(t) = M(t) - \lambda t \tag{2.33}$$

is a local martingale of zero mean under the family of sigma algebras $\{\mathcal{F}_t\}$; this is seen by computing $E[M(t)-M(s)|\mathcal{F}_s]$ from (2.27). According to the optional sampling theorem ([15], Theorem VI.T13), the martingale relation remains valid if ordinary times are replaced by stopping times respective to $\mathcal{F}_t$. C is such a stopping time, and M(C) = N. Hence, the evaluation of E[K(C)] = 0 yields (2.32) directly.

The results obtained thus far can be summarized to enunciate sufficient conditions under which Little's formula is valid for a Markov queueing system.

Theorem 2.14: Let Z be an irreducible positive recurrent Markov queueing process. Assume that customers enter the system singly, and that the entry process satisfies Hypothesis 2.10. Suppose further that the number N of customers in a (any) busy cycle is aperiodic, with $E(N^2) < \infty$. Then L(t) converges in distribution to a random variable whose expectation L is finite, where L(t) is the number of customers in the system at time t. Similarly, the sojourn times W_n of customers (ordered by their chronology of entry) converges in distribution to a variable of finite expectation W. These expectations are related through Little's formula,

$$L = \lambda W \quad , \tag{2.34}$$

in which λ is the expected rate of entry of customers.

III. TWO EXTENSIONS OF LITTLE'S FORMULA

In this Section we examine the effects of some relaxations of the hypotheses required for Theorem 2.14. The appropriate questions are: does Little's formula still hold, perhaps in modified form? If so, how must (2.34) be changed? The most likely candidates for less restrictive assumptions are those on the periodicity of N, and the state invariance (see Hypothesis 2.10) of the customer entry rate. Each of these is analyzed in turn.

We have already seen from Example 0.1 that N may be periodic; in fact, the periodicity of N is trivial if we permit only batch service of a single specified size. However, N can also be periodic in simple inifinite capacity systems with unit arrivals and service, as is indicated by

Example 3.1: The periodicity of N depends only on the set of possible transitions of the discrete process $\{Z^n\}$, so that we must specify which of the q_{xz} (see (2.2)) are zero. The other q_{xz} and the $\alpha(z)$ are arbitrary, subject to $\sum_z q_{xz} = 1$ and the positive recurrence of Z. With this in mind, we concentrate on specifying allowable one-step transitions of $\{Z^n\}$. Specifically, we construct these transitions to give N a periodicity of two, although the same technique can be used to yield other periodicities for N.

The state space $\mathcal{X}$ of Z is partitioned into subsets $\mathcal{X}_k = \{z: ||z|| = k\}$. Let $J(0) = \theta$, and each $\mathcal{X}_k$, $k \geq 1$, is the union of

disjoint nonempty subsets I(k) and J(k). The following one-step transitions are the only ones allowed: $J(0)\to I(1)$, and for each $k \geq 1$, $I(k)\to J(k+1)$, $J(k)\to I(k+1)$, and $J(k)\to J(k-1)$. If M^n is the number of customer arrivals in the first n epochs, M^n is even if $Z^n \varepsilon J(k)$ and odd if $Z^n \varepsilon I(k)$ (any k); this is readily shown by induction. Therefore, M^n is even whenever $Z^n = \theta$, and this reflects the periodicity of N.

For instance, we may take a system consisting of two disconnected exponential servers with state-dependent assignment of incoming customers, and service by only one server at each state. In terms of the I(k) and J(k), we let $J(2m) = (m,m)$, $J(2m+1) = (m,m+1)$, $I(2m) = (m-1,m+1)$, and $I(2m+1) = (m+1,m)$. As was noted earlier, service and customer accession rates may be chosen arbitrarily.

The effect of periodic M on Little's formula is minor, as is indicated by

Theorem 3.2: Let the number of customers in a busy cycle have periodicity d. Then the conclusions of Theorem 2.14 remain unchanged, except that the sojourn times W_n may fail to converge in distribution. Instead, as $n \to \infty$

$$d^{-1}\sum_{j=1}^{d} W_{nd+j} \to W_\infty \quad , \tag{3.1}$$

the convergence again being in distribution.

Proof: Theorem 3.2 in [6] is modified by insertion of the discrete renewal theory result ([5], p. 313) for periodic renewal processes. In other words, the basic formula $L = \lambda W$ remains

unchanged, with the meaning of W_∞ generalized by (3.1) to arbitrary $d \geq 1$.

We next turn to the consideration of arbitrary customer entry rates, which are assumed to depend on the state of the system, but not directly on time. We shall suppose (as before) that customers arrive singly, but that (cf. (2.26))

$$\lim_{h \searrow 0} h^{-1} P[M(s+h)-M(s)=1 \,|\, Z(s)=z] = \lambda_z \qquad (3.2)$$

need not be a constant. Call

$$\hat{\lambda} = \sup_{z \in X} \lambda_z \quad , \qquad (3.3)$$

and assume that $\hat{\lambda}$ is finite. We can then modify the system as follows: whenever $Z(t) = z$, there is a additional Poisson customer input stream $\tilde{M}$ of intensity $(\hat{\lambda}-\lambda_z)$, which is conditionally independent of M, given Z(t). Any arrival under $\tilde{M}$ passes through the system instantaneously, so that Z is unchanged, and the corresponding $W_n = 0$. The total arrival stream $(M+\tilde{M})$ then satisfies Hypothesis 2.10. To investigate the other hypotheses underlying Theorem 2.14, we first note that a rigorous construction of the modified process $\hat{Y}$ including the "passed through" customers is easily achievable. One augments Y by $\tilde{M}$, and adjusts U to properly index the arriving customers to include those attributable to $\tilde{M}$. We shall omit the details, which are tedious and routine. Let us denote the parameters of the modified process consistently by carats. Relations between the old and new processes include

$$\int_0^{\hat{C}} \hat{L}(t)dt = \int_0^{C} L(t)dt \qquad (3.4)$$

and

$$\sum_{n=1}^{\hat{N}} \hat{W}_n = \sum_{n=1}^{N} W_n \tag{3.5}$$

If $E(N^2) < \infty$, Theorem 2.9 assures that all the expectations of the terms in (3.4) and (3.5) are finite. Moreover, N is aperiodic if $\lambda_z < \lambda$ for at least one z. Theorem 2.14 then asserts that

$$\hat{L} = \hat{\lambda}\hat{W} \quad . \tag{3.6}$$

Since $\hat{L}(t) = L(t)$ for each t, $\hat{L}$ can be replaced by L in (3.6). A formula involving L and W can then be derived if W is related to $\hat{W}$.

Indeed, (1.3) is valid for both $\hat{W}$ and W, so that

$$\hat{W} = [E(N)]\{E(\hat{N})\}^{-1}W \tag{3.7}$$

in view of (3.5). To evaluate the ratio $E(N)/E(\hat{N})$, take N_z as the number of customers entering the original system during the busy cycle while $Z(t) = z$. Now

$$E[\hat{N}_z - N_z \,|\, Z(t)=z \text{ for duration } \tau_z] = (\hat{\lambda}-\lambda_z)\tau_z \tag{3.8}$$

and (see [5], p. 263)

$$E(\tau_z) = \pi_z E(C) \quad , \tag{3.9}$$

in which π_z is the equilibrium probability $\pi_z = \lim_{t\to\infty} P[Z(t)=z]$. The resulting $E(\hat{N}_z - N_z) = E(C)\{(\hat{\lambda}-\lambda_z)\pi_z\}$ can be summed to obtain $E(\hat{N}-N)$. Moreover, Theorem 2.13 applies to the modified system, so that $E(\hat{C}) = \hat{\lambda}^{-1}E(\hat{N})$, with $E(\hat{C}) = E(C)$ because $\hat{C} = C$. It follows that

$$E(N) = E(C)\{\sum_{z\in X} \lambda_z \pi_z\} \tag{3.10}$$

which is itself a useful formula, and

$$E(N)\{E(\hat{N})\}^{-1} = \hat{\lambda}^{-1}\{\sum_z \lambda_z \pi_z\} \quad . \tag{3.11}$$

By virtue of (3.11), we can claim

Theorem 3.2: Let Z be an irreducible positive recurrent Markov queueing process with single customer entry at rates λ_z specified by (3.2). If $E(N^2) < \infty$, L(t) converges in distribution to a random variable whose expectation L is finite, and the W_n converge in the sense of Theorem 3.1 to a random variable of expectation W. L and W are related through Little's formula, which takes the form

$$L = \bar{\lambda} W \quad . \tag{3.12}$$

Here $\bar{\lambda}$ is defined as $\bar{\lambda} = (\sum_z \lambda_z \pi_z)$.

Evidently, $\bar{\lambda}$ can be identified with the mean customer entrance rate, since π_z is also the average proportion of time the system spends in state z. Indeed, $\bar{\lambda}$ has the further interpretation as the almost sure limit of $t^{-1}M(t)$, so that $\bar{\lambda}$ appears when Little's formula is taken in the time average sense (cf. [18]).

VI. PRIORITY INTERRUPTS AND CLASSES OF CUSTOMERS

To broaden the applicability of Little's relation further, we extend our model to admit network behaviors discussed in recent (computer oriented) papers [1][3][13]. We shall begin with a consideration of priority interrupt queueing disciplines, in which a customer currently in service is pushed aside in favor of a new arrival; once the new customer completes service at the station, the order of resuming incompleted services is prescribed by the queueing discipline.

Because of the "forgetting" property of the exponential distribution, an interrupt at an exponential server is equivalent to a mere reordering of the customers at the station. This is easily provided by the specification of $\{U^n\}$. For a service distribution whose Laplace transform is rational, the approach is more complex. Ordinarily, the single station is represented as m substations, where m is the degree of the denominator of the rational transform [7][17]. Appropriate blocking (no service provided at more than one substation at a time) then leads to the correct model. This structure can be extended to priority interrupts by paralleling all but the first substation by "dummy substations" incapable of providing service. Upon arrival of a new customer, the customer (if any) at substation k is "laid aside" to the corresponding dummy station. When the new customer has passed through all m substations, one of the customers who has been laid aside is selected by the discipline to continue service by returning immediately to the real substation from whence he came. Since all substations function as exponential servers, the above technique becomes valid by virtue of the "forgetting" property at each substation.

The construction of the preceding paragraph leaves untouched the consistency of $\{U^n\}$ and $\{Z^n\}$, so that Theorem 2.1 holds as before. Thus (cf. Theorem 2.14), Little's formula holds for priority interrupt systems under the same hypotheses as for any of the other disciplines mentioned. Naturally, the extensions of Section III will apply to priority interrupt queues also.

We now turn to systems with multiple customer types. Several versions of Markov queueing systems with different classes of

customers have been proposed [1][13], but not all of these are consistent with Little's result for the individual classes. In particular, transmutation of customers from one class to another can invalidate Little's formula. A simple example is adduced by considering tandem exponential servers, with entrance of class one customers only to the first server. As each customer passes on to the second server, he becomes a class two customer, and eventually departs as such. A direct calculation shows that Little's equality holds for neither customer class. To avoid such difficulties, our model maintains the integrity of the class of each customer as he passes through the network. However, we do not restrict priority assignments, or service rates and routings depending on the number and types of customers at each service station.

A prefix is used to denote each of the c customer classes. Specifically, the line length process at station j takes on the vector form $Z_j = ({}_1Z_j, {}_2Z_j, \ldots {}_cZ_j)$. Similarly, ${}_i\lambda$ is the rate of arrival for customers of class i, ${}_iM$ is the corresponding arrival process, ${}_iW_n$ the waiting time for the n'th customer of class i, and so forth. It is necessary to expand $\{U^n\}$ so that U^n has entries (m, i_m, a_m, q_m, j_m). In this quintuple, m is the customer's order of arrival in chronological order, i_m is his class, a_m is the customer's order of arrival among customers of class i_m, q_m the queueing station to which customer m is assigned, and j_m his position in the queue at q_m. As before, $q_m = j_m = 0$ for a customer who has already exited the system.

We would anticipate that Little's formula appears as

$$ {}_iL = ({}_i\lambda)\,{}_iW \quad ; \tag{4.1} $$

such is indeed the case. In the first place, Theorem 2.1 now takes the form

$$ \sum_{n=1}^{{}_iN} {}_iW_n = \int_0^C {}_iL(t)\,dt \quad . \tag{4.2} $$

Next, as in Theorem 2.4, each $\{{}_iW_n\}$ is regenerative with respect to the renewal sequence $\{{}_iN_k\}$. If ${}_iN$ is aperiodic and if $E(N^2)$, we can claim

$$ {}_iW = \{E({}_iN)\}^{-1} E\{\sum_{n=1}^{{}_iN} {}_iW_n\} \tag{4.3} $$

and (cf. Theorem 2.9)

$$ {}_iL = \{E(C)\}^{-1} E\{\int_0^C {}_iL(t)\,dt\} \quad . \tag{4.4} $$

To complete the derivation of (4.1), we must relate $E({}_iN)$ to $E(C)$. For this purpose, let us suppose that Hypothesis 2.10 holds separately for each ${}_iM$. The martingale decomposition of Theorem 2.13, applied to ${}_iM$ at the stopping time C, yields

$$ E({}_iN) = ({}_i\lambda)E(C) \quad . \tag{4.5} $$

Finally, (4.1) results from combining (4.2) through (4.5). As in the case of one type of customer, (4.1) can be generalized to include periodic ${}_iN$ and state-dependent customer entry rates.

V. NON-MARKOVIAN SYSTEMS

The Markov nature of the queueing process Z is by no means essential to the derivation of Little's formula; for instance, the formula applies to the stable G/G/s queue [17], which is certainly not Markov. However, any assertions on Little's relation pertaining to queueing networks with arbitrary service and entry characteristics is necessarily limited to generalities. We can say that if Z satisfies the regenerative properties (2.11) and (2.12), meets the finiteness conditions $E(N) < \infty$ and (2.23), and has a Poisson process exogenous input, the standard arguments yield $L = \lambda W$.

When there is only one mode of entrance to the system, these regeneration points can be taken as the starting instants of the busy period; this suggests a new freedom in the choice of arrival process. In fact, the length of the (first) busy cycle is now provided by (1.4). That Z is regenerative respective to $\{C_k\}$ carries the implication that the interarrival process $\{A_k\}$ is regenerative with respect to numbers of customers $\{N_k\}$ in the respective busy cycles. If, further, N is aperiodic and $E(N) < \infty$, the A_n converge in distribution to a random variable A whose expectation is

$$E(A) = \{E(N)\}^{-1} E\{\sum_{n=1}^{N} A_n\} \quad . \tag{5.1}$$

We see that (5.1) implies (1.5), but without any requirements that the A_n be i.i.d. random variables. Finally, if we take $\lambda = [E(A)]^{-1}$, we recover Little's formula $L = \lambda W$.

An attempt to impose a more specific probability structure on Z seems to be less successful. For instance, we find that a semi-Markov process Z is also regenerative respective to $\{C_k\}$ (cf. [5], Section 10.6), but are unable to turn this observation to advantage in generating a broader class of queueing network models. We are also aware of the feasibility of approximating general service time distributions by Erlang service in the context of Markov systems (see [17] for a rigorous treatment); although this approach works for some aspects of the G/G/s queue [17], there seems little chance that the analysis can be extended to the approximation of L(t) and $\{W_n\}$ in queueing networks. Consequently, we have limited detailed consideration to Markov queueing systems, which are adequate to describe a rich variety of queueing system behaviors.

REFERENCES

[1] Baskett, F., Chandy, K. M., et al., "Open, closed, and mixed networks of queues with different classes of customers," J. ACM, 22(1975), 248-260.

[2] Beutler, F. J., Melamed, B., and Zeigler, B. P., "Equilibrium properties of arbitrarily interconnected queueing networks," in Multivariate Analysis IV, P. R. Krishnaiah (ed.), North-Holland, Amsterdam, 1977.

[3] Chandy, K. M., Howard Hr., J. H., Towsley, D. F., "Product form and local balance in queueing networks," J. ACM, 24 (1977), 250-263.

[4] Chung, K. L., Markov Chains, 2nd ed., Springer, New York, 1967.

[5] Cinlar, E., Introduction to Stochastic Processes, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[6] Cohen, J. W., On Regenerative Processes in Queueing Theory, Springer-Verlag, New York, 1976.

[7] Cox, D. R., and Smith, W. L., Queues, Chapman and Hall London, 1961.

[8] Feller, W., An Introduction to Probability Theory and Its Applications, vol. I, 3rd ed., Wiley, New York, 1968.

[9] Gelenbe, E., Muntz, R. R., "Probabilitistic models of computer systems--part I (exact results)," Acta Informatica, 7(1976) 35-60.

[10] Hille, E., Analytic Function Theory, Vol. I, Ginn and Co. Boston, 1959.

[11] Jackson, J. R., "Jobshop-like queueing systems," Management Sci., 10(1963), 131-142

[12] Jewell, W. S., "A simple proof of: $L = \lambda W$," Operations Research, 15(1967), 1109-1116.

[13] Kelly, F. P., "Networks of queues with customers of different types," J. Appl. Prob., 12(1975), 542-554.

[14] Kobayashi, H., and Konheim, A. G., "Queueing models for computer communications system analysis," IEEE Trans. Comm., COM-25(1977), 2-29.

[15] Meyer, P., Probability and Potentials, Blaisdell, Waltham, MA, 1966.

[16] Muntz, R. R., and Baskett, F., "Open, Closed and Mixed Networks of Queues with Different Classes of Customers," Digital Systems Laboratory, Stanford University, 1972.

[17] Schassberger, R., Warteschlangen, Springer-Verlag, New York, 1973.

[18] Stidham, Jr., S., "L = λW: a discounted analogue and a new proof," Operations Research, 20(1972), 1115-1126.

[19] Stidham, Jr., S., "A last word on L = λW," Operations Research, 22(1974), 417-421.